(51) International Patent Classification ^{6} : G10L 5/06	A1	(11) International Publication Number: **WO 99/54869**
		(43) International Publication Date: 28 October 1999 (28.10.99)

(54) Title: ADAPTATION OF A SPEECH RECOGNIZER FOR DIALECTAL AND LINGUISTIC DOMAIN VARIATIONS

(57) Abstract

The invention relates to a generator and a method for generating an adapted speaker independent speech recognizer. The generator of an adapted speech recognizer is being based upon a base speech recognizer of an arbitrary base language. The generator also comprises an additional speech data corpus used for generation of said adapted speech recognizer. Said additional speech data corpus comprises a collection of domain specific speech data and/or dialect specific speech data. Said generator comprises reestimation means for reestimating language or domain specific acoustic model parameters of the base speech recognizer by a speaker adaption technique. Said additional speech data corpus is exploited by said reestimation means for generating the adapted speech recognizer. The invention proposes smoothing means for smoothing the reestimated acoustic model parameters. A beneficial range for the smoothing factor of a Bayesian smoothing is given. It is further suggested to iterate the adaption process.

D E S C R I P T I O N

ADAPTATION OF A SPEECH RECOGNIZER FOR DIALECTAL AND LINGUISTIC DOMAIN VARIATIONS

1 Background of the Invention

1.1 Field of the Invention

The present invention relates to speech recognition systems. More particularly, the invention relates to a generator for generating an adapted speech recognizer. Furthermore the invention also relates to a method of generating such an adapted speech recognizer said method being executed by said generator.

1.2 Description and Disadvantages of Prior Art

For more than two decades speech recognition systems use Hidden Markov Models to capture the statistical properties of acoustic subword units, like e.g. context dependent phones or subphones. An overview on this topic may be found for instance in L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77(2), pp. 257-285, 1989 or in X. Huang and Y. Ariki and M. Jack, Hidden Markov Models for Speech Recognition, Information Technology Series, Edinburgh University Press, Edinburgh, 1990.

A Hidden Markov Model is a stochastic automaton that operates on a finite set of states $S = \{s_1, \ldots, s_N\}$ and allows for the observation of an output each time $t, t = 1, 2, \ldots, T$ a state is occupied. It is defined by a tuple $HMM = (\pi, A, B)$ where the initial state vector

$$\pi = [\pi_i] = [P(s(1) = s_i)], \quad 1 \leq i \leq N, \quad (1)$$

gives the probabilities that the HMM occupies state s_i at time $t = 1$, and

$$A = [a_{ij}] = [P(s(t+1) = s_j \mid s(t) = s_i)], \quad 1 \quad i, j \quad N, \qquad (2)$$

gives the probabilities for a transition from state s_i to s_j, assuming a first order time invariant process. In case of discrete HMMs the observations o_l are from a finite alphabet $O = \{o_1, \ldots, o_L\}$, and

$$B = [b_{kl}] = [p(o_l \mid s(t) = s_k)], \quad 1 \le k \le N, 1 \quad l \quad L, \qquad (3)$$

is a stochastic matrix that gives the probabilities to observe o_l in state s_k.

For (semi-)continuous HMMs, which provide the state of the art in today's large vocabulary continuous speech recognition systems, the observations are (continuous valued) feature vectors c, and the output probabilities are defined by the probability density functions

$$B = [b_{kl}] = [p(c_l \mid s(t) = s_k)], \quad 1 \le k \le N, 1 \quad l \quad L, \qquad (4)$$

The actual distribution $p(c_l \mid s_k)$ of the feature vectors is usually approximated by a mixture of N_k Gaussians:

$$p(c_l \mid s_k) = \sum_{i=1}^{N_k} \omega_{ik} N(c_l \mid \mu_{ik}, \Sigma_{ik}) \qquad (5)$$

$$= \sum_{i=1}^{N_k} \omega_{ik} \cdot |2\pi\Sigma_{ik}|^{-1/2} \cdot \exp(-(c_l - \mu_{ik})_T \Sigma_{ik}^{-1}(c_l \quad \mu_{ik})/2); \qquad (6)$$

the mixture component weights ω, the means μ, and the

covariance matrices Σ are estimated from a large amount of

transcribed speech data during the training of the recognizer.
A well known procedure to solve that problem is the EM-algorithm
(illustrated for instance by A. Dempster and N. Laird and D.
Rubin, Maximum Likelihood from Incomplete Data via the EM
Algorithm, Journal of the Royal Statistical Society, Series B
(Methodological), 1977, Vol. 39(1), pp. 1-38), and the markov

model parameters π, A, B are usually estimated by the use of the

forward-backward algorithm (illustrated for instance by L.
Rabiner, A Tutorial on Hidden Markov Models and Selected
Applications in Speech Recognition, Proceedings of the IEEE,
Vol. 77(2), pp. 257-285, 1989).

The training of a speech recognizer for an arbitrary language is
described in some detail by L. Bahl and S. Balakrishnan-Aiyer
and J. Bellegarda and M. Franz and P. Gopalakrishnan and D.
Nahamoo and M. Novak and M. Padmanabhan and M. Picheny and S.
Roukos, Performance of the IBM large vocabulary continuous
speech recognition system on the ARPA Wall Street Journal task,
Detroit, Proc. of the IEEE Int. Conference on Acoustics, Speech,
and Signal Processing, pp. 41-44, 1995 or by L. Bahl and P. de
Souza and P. Gopalakrishnan and D. Nahamoo and M. Picheny,
Context-dependent Vector Quantization for Continuous Speech
Recognition, Minneapolis, Proc. of the IEEE Int. Conference on
Acoustics, Speech, and Signal Processing, 1993. The procedure is
briefly outlined in the following, since it provides the basis
for the current invention. The algorithm assumes the existence
of a labelled training corpus and a speaker independent
recognizer for the computation of an initial alignment between
the spoken words and the speech signal. After the framewise
computation of cepstral features and their first and second
order derivatives, the Viterbi algorithm is used for the
selection of phonetic baseforms that best matches the
utterances. An outline of the Viterbi algorithm can be found in

Viterbi, A.J., Error Bounds for Convolutional Codes and an
asymptotically optimum Decoding Algorithm, IEEE Trans. on
Information Theory, Vol. 13, pp. 260--269, 1967.

Since the acoustic feature vectors show significant variations
in different contexts, it is important to identify the phonetic
contexts that lead to specific variations. For that purpose the
labelled training data is passed through a binary decision
network that separates the contexts into equivalence classes
depending on the variations observed in the feature vectors. A
multi-dimensional Gaussian mixture model is used to model the
feature vectors that belong to each class represented by the
terminal nodes (leaves) of the decision network. These models
are used as initial observation densities in a set of
context-dependent, continuous parameter HMM, and are further
refined by running the forward-backward algorithm, which
converges to a local optimum after a few iterations. The total
number of both context dependent HMMs and Gaussians is limited
by the specification of an upper bound and depends on the amount
and contents of the training data

Both the large amount of data needed for the estimation of model
parameters and relevant contexts and the need to run several
forward-backward iterations make the training of a speech
recognizer a very time consuming process. Moreover, speakers
have to face a large degradation in recognition accuracy, if
their pronunciation differs from those observed during the
training of the recognizer. This can be caused by poorly trained
acoustic models due to a mismatch between the collected data and
the task domain. This can be considered as the main reason for
the fact that most commercially available speech recognition
products (like e.g. IBM ViaVoice, Dragon Naturally Speaking,
Kurzweill) at least recommend, if not enforce, a new user to
read an enrollment script of about 50 - 250 sentences for a
speaker dependent reestimation of the model parameters.

For such reestimation processes for instance speaker adaptation
techniques like the **maximum a posteriori** estimation of gaussian
mixture observations (MAP adaptation) - refer for instance to J.
Gauvain and C. Lee, Maximum a Posteriori Estimation of
Multivariate Gaussian Mixture Observations of Markov Chains,
IEEE Trans. on Speech and Audio Processing, Vol. 2(2), pp.
291--298, 1994 - or the **maximum likelihood linear regression**
(MLLR adaptation) - refer for instance to C. Leggetter and P.
Woodland, Maximum likelihood linear regression for speaker
adaptation of continuous density hidden Markov models, Computer
Speech and Language, Vol. 9, pp. 171--185, 1995 - are exploited
during the training of the recognizer.

1.3 Objective of the Invention

Given these problems, the invention is based on the objective of
a reduction of training effort for individual end users and an
improved speaker independent recognition accuracy.

It is a further objective of the current invention to improve
the easiness and the rapidness of development of new adapted
speech recognizers.

2 Summary and Advantages of the Invention

The objective of the invention is solved by the independent
claims.

The objective of the invention is solved by claim 1.
The generator of an adapted speech recognizer according the
teaching of the current application is being based upon a base
speech recognizer 201 for a definite but arbitrary base
language. The generator also comprises an additional speech data
corpus 202 used for generation of said adapted speech
recognizer. Said additional speech data corpus comprises a
collection of domain specific speech data and/or dialect

specific speech data. Furthermore said generator comprises reestimation means 203 for reestimating acoustic model parameters of the base speech recognizer by a speaker adaption technique. Said additional speech data corpus is exploited by said reestimation means for generating the adapted speech recognizer.

The technique proposed by the current invention thus achieves a significant reduction of training effort for individual end users, an improved speaker independent recognition accuracy for specific domains and dialect speakers, and the rapid development of new data files for speech recognizers in specific environments. Moreover also the recognition rate of non-dialect speakers is also improved.

Whereas in the past speaker adaptation techniques were usually applied to an individual end users speech data and therefore yield in a speaker dependent speech recognizer, in the current invention they are applied to a dialect and/or domain specific collection of training data from several speakers. This allows for an improved speaker independent recognition especially (but not solely) for a given dialect and domain and minimizes the individual end users investment to customize the recognizer to their needs.

Another important aspect of this invention is the reduced effort for the generation of a specific speech recognizer: whereas other commercially available toolkits start from the definition of subword units and/or HMM topologies, and thus require a considerable large amount of training data, the current approach starts from an already trained general purpose speech recognizer.

The approach of the current teaching offers a scalable recognition accuracy, if dialects and/or specific domains are handled in an integrated speech recognizer. As the current invention is completely independent from the specific dialect

- 7 -

and/or specific domain they may be combined in any possible combination.

Moreover the amount of additional data (the additional speech data corpus) is very moderate. Only few additional domain specific or dialect data is required and besides that it is inexpensive and easy to collect.

Finally the current invention allows to reduce the time for the upfront training of the recognizer significantly. Therefore it allows for rapid development of new data files for recognizers in specific environments or combination of environments.

Additional advantages are accomplished by claim 2.
According to a further embodiment of the proposed invention said additional speech data corpus can be collected unsupervised or supervised.

Based on such a teaching complete flexibility is offered to an exploiter of the current teaching on how the additional speech data corpus is being provided.

Additional advantages are accomplished by claim 3.
According to a further embodiment of the proposed invention said acoustic model is a Hidden-Markov-Model (HMM).

Thus the current teaching my be applied to the HMM technology. Therefore the HMM approach, one of the most successful techniques in the area of speech recognition, can be further improved with the current teachings.

Additional advantages are accomplished by claim 4.
According to a further embodiment of the proposed invention said speaker adaption technique is the Maximum-A-Posteriori-adaption (MAP) or the Maximum-Likelihood-Linear-Regression-adaption (MLLR).

These approaches allow also to deal with situations in which
only sparse training data is available. Excellent adaptation
results in terms of recognition accuracy and generation speed of
the adapted speech recognizer are achieved with especially these
speaker adaptation techniques.

Claim 5 achieves additional benefits.
According to this additional embodiment of the proposed
invention smoothing means 204 are introduced for optionally
smoothing the reestimated acoustic acoustic model parameters.

Experiments revealed that additional smoothing further improves
the recognition accuracy and the adaptation speed. Especially in
cases with a limited amount of training data these improvements
are of specific importance.

Additional advantages are accomplished by claim 6, 7 and 8.
According to a further embodiment of the proposed invention said
smoothing means performing a Bayesian smoothing. A smoothing
factor K from the range 1 to 500 is being suggested. Especially
the subrange for smoothing factor K of 20 to 60 is proposed.

Bayesian smoothing has been shown to produce good results in
terms of recognition accuracy and performance. Intensive
experimentation revealed that a smoothing factor K from the
range 1 to 500 accomplishes good results. Especially the
subrange for smoothing factor K of 20 to 60 turned out to
achieve the best results.

Additional advantages are accomplished by claim 9.
According to a further embodiment of the proposed invention
iteration means 205 for optionally iterating the operation of
said reestimation means and for optionally iterating the
operation of said smoothing means are suggested. The iteration
may be based on said reestimated dialect or domain specific
acoustic model parameters or based on said base language
acoustic model parameters.

This teaching allows for a stepwise approach to the generation
of an optimally adapted speech recognizer.

Additional advantages are accomplished by claim 10.
According to a further embodiment of the proposed invention said
iteration means use a modified additional speech data corpus
and/or said iteration means use a new smoothing factor value K.

With this teaching a remarkable amount of selective influence on
the iteration process is possible. Depending on the nature of
said additional speech data corpus the iteration process may be
based on an enlarged or modified additional speech data corpus.
For instance a changed smoothing factor allows to assist the
generation process depending on the narrowness of the amount of
training data.

Additional advantages are accomplished by claim 11.
According to a further embodiment of the proposed invention said
adapted speech recognizer is speaker independent.

This approach offers at the same time the benefit that an
adapted speech recognizer can be generated which is already
tailored to a certain domain and/or dialect or set of domains
and/or dialects but which still is speaker independent.
Nevertheless said adapted speech recognizer may be further
personalized resulting in a speaker dependent speech recognizer.
Thus at the same time specialization and flexibility is offered
at the same time.

The objective of the invention is solved by claim 12.
A method for generating an adapted speech recognizer using a
base speech recognizer 201 for a definite but arbitrary base
language is suggested. Said method comprises a first step 202 of
providing an additional speech data corpus. Said additional
speech data corpus comprises a collection of domain specific

speech data and/or dialect specific speech data. Furthermore said method comprises a second step 203 of reestimating acoustic model parameters of said base speech recognizer by a speaker adaption technique using said additional speech data corpus.

The benefits achieved by teaching of claim 12 are those already discussed with claim 1.

Additional advantages are accomplished by claim 13.
According to a further embodiment of the proposed invention said method comprises an optional third step 204 for smoothing the reestimated acoustic model parameters.

Experiments revealed that additional smoothing further improves the recognition accuracy and the adaptation speed. Especially in cases with a limited amount of training data these improvements are of specific importance. For further advantages refer to the benefits discussed with claim 6, 7, and 8 above.

Additional advantages are accomplished by claim 14.
According to a further embodiment of the proposed invention said method comprises an optional fourth step 205 for iterating said first step by providing a modified additional speech data corpus and for iterating said second and third step based on said reestimated acoustic model parameters or based on said base acoustic model parameters.

For advantages adhering to this teaching refer to the benefits discussed with claim 9 above.

Additional advantages are accomplished by claim 15.
According to a further embodiment of the proposed invention said acoustic model is a Hidden Markov Model (HMM). Moreover it is taught that said speaker adaption technique is the Maximum-A-Posteriori-adaption (MAP) or the Maximum-Likelihood-Linear-Regression-adaption (MLLR). In addition it is suggested to perform a Bayesian smoothing.

The advantages of this approach has been discussed with claim 3,
4 and claims 6, 7 and 8 above.

Additional advantages are accomplished by claim 16.
According to a further embodiment of the proposed invention said
adapted speech recognizer is speaker independent.

Benefits related to this teaching are discussed together with
claim 11 above.

3 Brief Description of the Drawings

Figure 1   is a diagram reflecting the overall structure of the
            state-of-the-art adaptation process visualizing the
            generation of a speaker dependent speech recognizer
            from a speaker independent speech recognizer of the
            base language.
Figure 2   is a diagram reflecting the overall structure of the
            adaptation process according the current invention
            visualizing the generation of an improved speaker
            independent speech recognizer from a speaker
            independent speech recognizer of the base language.
            Said improved speaker independent speech recognizer
            may be the basis for further customization generating
            an improved speaker dependent speech recognizer.
Figure 3   gives a comparison of the error rates of the baseline
            recognizer (VV), the standard training procedure
            (VV-S), and the scalable fastboot method (VV-G)
            normalized to the error rate of the baseline
            recognizer (VV) for a German test speaker.

4 Description of the Preferred Embodiment

Throughout this description the usage of current teaching is not
limited to a certain language, a certain dialect or a certain
usage domain. If a certain language, a certain dialect or a

certain domain is mentioned this is to be interpreted as an example only not limiting the scope of the invention.

Moreover if the current description is referring to a dialect/domain this may be interpreted as a specific dialect/domain or any combination of dialects/domains.

4.1 Introduction

The training of a for instance Hidden Markov Model based speech recognizer for a given language requires the collection of a large amount of general speech data for the detection of relevant phonetic contexts and the proper estimation of acoustic model parameters. However, a significant decrease in recognition accuracy can be observed, if a speaker's pronunciation differs significantly from those present in the training corpus, Therefore, commercially available speech recognizers partly impose the estimation of acoustic parameters to the individual end-user, by enforcing the personalization process depicted in Fig. 1.

The starting point is a speech recognizer 101 for a base language which is speaker independent and without specialization to any domain. The individual user has to read a predefined enrollment script 103 which is a further input to the reestimation process 102. Within this reestimation process the parameters of the underlying acoustic model are adapted by available speaker adaptation techniques according to the state of the art. The result of this generation process is the output of a speaker dependent speech recognizer.

The current invention is teaching a fast bootstrap (i.e. upfront) procedure for the training of a speech recognizer with improved recognition accuracy; i.e. the current invention is proposing a generation process for an additionally adapted speaker independent speech recognizer based upon a general speech recognizer for the base language.

According to the current teaching both accuracy and speed of the
recognition system can be significantly improved by explicit
modelling of language dialects and orthogonally by the
integration of domain specific training data in the modelling
process. The architecture of the invention allows to improve the
recognition system along both of these directions. The current
invention utilizes the fact that for certain dialects, like e.g.
Austrian German or Canadian French, the phonetic contexts are
similar in the base language (German or French, resp.), whereas
acoustic model parameters differ significantly due to different
pronunciations. Similar, not well trained acoustic models for
specific domains (e.g. base domain: office correspondence,
specific domain: radiology) can be estimated more accurate by
the application of the invention to a limited amount of acoustic
data from the target domain.

By upfront training of dialects and/or specific domains towards
a large number of end users the performance of the recognition
system is tremendously increased and user investment to
customize the recognizer to their needs is minimized.

According the current teaching it is in addition possible to
reduce the training procedure to the computation of Hidden
Markov Model parameters. Moreover, it is possible to use
Bayesian smoothing techniques for the better utilization of a
small amount of dialect or domain specific training data and for
the achievement of a scalable recognition accuracy for a
specific dialect within a base language (or domain, resp.).

Thus, based on these techniques, the current invention achieves
the reduction of training efforts for individual end users, an
improved speaker independent recognition accuracy for specific
domains and dialect speakers, and the rapid development of new
data files for speech recognizers in specific environments.

4.2 Solution

The current invention (called **fastboot** in the remainder)
utilizes the observation that speaker adaptation techniques,
like e.g. the maximum a posteriori estimation of gaussian
mixture observations (MAP adaptation) or maximum likelihood
linear regression (MLLR adaptation), yield a significant larger
improvement in recognition accuracy for dialect speakers than
for speakers that use pronunciations observed during the
training of the recognizer. According to the current teaching
this approach results in improved speaker independent
recognition accuracy not only for dialect speakers. These
techniques move the output probabilities B of the HMMs to a
speakers particular acoustic space, and thus it is achieved that

    o   the main differences between dialect and base language
        are captured by the output probabilities of the HMMs,
    o   the trained parameters for the base language already
        provide good initial values for a dialect specific
        reestimation by the forward-backward algorithm, and
    o   the reestimation of significant contexts from dialect
        data can be omitted to achieve a fast training
        procedure.

The basic teaching of the current invention is depicted in Fig.
2, teaching the application of additional speaker adaptation
techniques for the upfront training, i.e. for the training
before the speech recognizer is personalized to a specific user,
of a speech recognizer for a dialect within a base language or
for a special domain.

Referring to Fig. 2 the current invention suggest to start with
base speech recognizer 201 for a base language. For the final
generation of an adapted speech recognizer an additional speech
data corpus 202 is being provided; the current invention is
suggesting the usage of actual speech data not comparable with a
dictionary. This additional speech data corpus may comprise any
collection of domain specific speech data and/or dialect

specific speech data. The speech recognizer for the base
language may be already used for an unsupervised collection of
the additional speech data.

The generation process comprises reestimating 203 the acoustic
model parameters of said base speech recognizer by one of the
available speaker adaption techniques using the additional
speech data corpus, thus generating an improved adapted speech
recognizer reducing the potential training effort for individual
end users and at the same time improving the speaker independent
recognition accuracy for specific domains and/or dialect
speakers.

Optionally the invention teaches the application of a further
smoothing 204 of the reestimated acoustic model parameters.
Bayesian smoothing is an efficient smoothing technology for that
purpose. With respect to Bayesian smoothing good results have
been achieved with a smoothing factor k from the range 1 to 500
(see below for more details with respect to the smoothing
approach). Especially the range of 20 to 60 for the smoothing
factor k ensued excellent results.

Optionally the current teaching suggests to iterate 205 the
above mentioned generation process of reestimating the acoustic
model parameters and the smoothing. The iteration can be based
on the reestimated acoustic model parameters of the previous run
or on the base acoustic model parameters. The iteration can be
based on the decision whether the generated adapted speech
recognizer shows sufficient recognition improvement. To achieve
the desired recognition improvements the iteration step may be
based for example on a modified additional speech data corpus
and/or on the usage of a new smoothing factor value K.

Finally the process results in the generation 206 of a adapted
speaker independent speech recognizer for dialect and/or
specific domain.

Whereas in the past speaker adaptation techniques were usually
applied to an individual end users speech data and therefore
yield in a speaker dependent speech recognizer, in the current
invention they are applied to a dialect and/or domain specific
collection of training data from several speakers. This allows
for an improved speaker independent recognition especially (but
not solely) for a given dialect and domain and minimizes the
individual end users investment to customize the recognizer to
their needs.

Another important aspect of this invention is the reduced effort
for the generation of a specific speech recognizer: whereas
other commercially available toolkits start from the definition
of subword units and/or HMM topologies, and thus require a
considerable large amount of training data, the current approach
starts from an already trained general purpose speech
recognizer.

For further recognition improvement this invention suggest to
optionally apply Bayesian smoothing to the reestimated

parameters. In particular it is suggested to use the means μ_i^b,

variances Γ_i^b and mixture component weights ω_i^b of the base

language system (distinguished by the upper index b) for the

reestimation of the dialect specific parameters μ_i^d, Γ_i^d and

ω_i^d (distinguished by the upper index d) by Bayesian smoothing

and tying (refer for instance to a J. Gauvain and C. Lee,
Maximum a Posteriori Estimation of Multivariate Gaussian Mixture
Observations of Markov Chains, IEEE Trans. on Speech and Audio
Processing, Vol. 2(2), pp. 291--298, 1994) according to the
following equations:

$$\mu_i^d = \frac{\sum_t c_i(t) x_t + \alpha_i \mu_i^b}{c_i + \alpha_i} \tag{7}$$

$$\Gamma_i^d = \frac{Y_i + \alpha_i (\Gamma_i^b + \mu_i^b \mu_i^{b,T})}{c_i + \alpha_i} - \mu_i^d \mu_i^{d,T} \tag{8}$$

$$Y_i = \sum_t c_i(t) x_t x_t^T \tag{9}$$

$$\omega_i^d = \frac{c_i + \alpha_i}{\sum_{m \in M} (c_m + \alpha_m)} , \qquad \alpha_j = k \cdot \omega_j^b \tag{10}$$

Here, $c_i = \sum_t c_i(t)$ is the sum of all posteriori probabilities

$c_i(t)$ of the i-th gaussian, at time t, computed from all

observed dialect data x_t, N denotes the total number of mixture

components, and M is the set of gaussians that belong to the

same phonetic context as the i-th gaussian. The constant k is

referred to as a smoothing factor; it allows for an optimization

of the recognition accuracy and depends on the relative amount

of dialect training data.

4.3 Example of an Embodiment of the Invention

In 1997 IBM Speech Systems released ViaVoice, the first
continuous speech recognition software available in 6 different
languages. The German recognizer, for example, was trained with
several hundred hours of carefully read continuous sentences.
Speech was collected solely from less than thousand native
German speakers (approx. 50 \% female, 50 \% male).

For test purposes of the current teaching 20 different German
speakers (10 female, 10 male) and 20 native Austrian speakers
(10 female, 10 male) were collected. All speakers read the same
medium perplexity test script from an office correspondence

domain, which is supposed to be one of the most important
applications for continuous speech recognition.

For both groups of speakers, Figure 3 compares the relative
speaker independent error rates achieved with the baseline
recognizer. Figure 3 shows a comparison of the error rates of
the baseline recognizer (VV), the standard training procedure
(VV-S), and the scalable fastboot method (VV-G) normalized to
the error rate of the baseline recognizer (VV) for the German
test speakers. The error rate for the Austrian speakers
increases by more than 50 percent, showing the need to improve
the recognition accuracy for dialect speakers. Therefore, for
the follow up product, ViaVoice Gold (VV-G), only less than 50
hours of speech from approx. hundred native Austrian speakers
(50 \% female, 50 \% male) have been collected and applied with
the fastboot approach for the upfront training of the recognizer
according to the current invention. Figure 3 compares the
results achieved with the fastboot method (VV-G) to the standard
training procedure (VV-S), that can be applied if both training
corpora are pooled together. It becomes evident that the
fastboot method is superior to the standard procedure and yields
a 30 percent improvement for the dialect speakers. The results
for different values of the smoothing factor show that
recognition accuracy is scalable, which is an important feature,
if an integrated recognizer for base language and dialect (or -
orthogonal to this direction - base domain and specific domain)
is needed. Moreover, since the pooled training corpus of the
common recognizer (VV-S) is approx. 7 times larger than the
Austrian training corpus, and usually the standard training
procedure has to compute 4 - 5 forward-backward iterations, the
fastboot method is at least 25 times faster. Thus, the rapid
development of speech recognizers for specific dialects or
domains becomes possible by our invention.

4.4 Further Advantages of the Current Teaching

The invention and its embodiment presented above demonstrate the following further advantages:

- The fastboot approach yields a significant decrease in speaker independent error rate for dialect speakers. Moreover also the recognition rate of non-dialect speakers is improved.
- The fastboot approach offers a scalable recognition accuracy, if dialects and/or specific domains are handled in an integrated speech recognizer.
- The fastboot approach uses only few additional domain specific or dialect data which is inexpensive and easy to collect.
- The fastboot approach reduces the time for the upfront training of the recognizer, and therefore allows for the rapid development of new data files for recognizers in specific environments.

5 Acronyms

HMM Hidden Markov Model
MAP maximum a posteriori adaptation
MLLR maximum likelihood linear regression adaptation

C L A I M S

1.   Generator of an adapted-speech-recognizer comprising a
        base-speech-recognizer (201) for a base-language

        further characterized by

        comprising an additional-speech-data-corpus (202) used
        for generation of said adapted-speech-recognizer and
        said additional-speech-data-corpus comprising a
        collection of domain-specific-speech-data and/or
        dialect-specific-speech-data, and

        said generator comprising reestimation-means (203) for
        reestimating acoustic-model-parameters of said base-
        speech-recognizer by a speaker-adaption-technique
        using said additional-speech-data-corpus.

2.   Generator according to claim 1

        wherein said additional-speech-data-corpus being
        provided by unsupervised or supervised collection.

3.   Generator according to any of above claims

        wherein said acoustic-model is a Hidden-Markov-Model
        (HMM).

4.   Generator according to claim 3

        wherein said speaker-adaption-technique is the
        Maximum-A-Posteriori-adaption (MAP) or

        wherein said speaker-adaption-technique is the
        Maximum-Likelihood-Linear-Regression-adaption (MLLR).

5.   Generator according to claim 4

        further comprising smoothing-means (204) for
        optionally smoothing the reestimated acoustic-model-
        parameters.

6.   Generator according to claim 5

        wherein said smoothing-means performing a Bayesian
        smoothing.

7.   Generator according to claim 6

        wherein a smoothing factor K is from the range 1 to
        500.

8.   Generator according to claim 6

        wherein a smoothing factor K is from the range 20 to
        60.

9.   Generator according to any of above claims

        further comprising iteration-means (205) for
        optionally iterating the operation of said
        reestimation-means and for optionally iterating the
        operation of said smoothing-means based on said
        reestimated-acoustic-model-parameters or based on said
        base-acoustic-model-parameters.

10. Generator according to claim 9

        wherein said iteration-means using a modified
        additional-speech-data-corpus and/or

        wherein said iteration-means using a new smoothing
        factor value K.

11. Generator according to any of above claims

    wherein said adapted-speech-recognizer being speaker
    independent.

12. Method for generating an adapted-speech-recognizer using a
     base-speech-recognizer (201) for a base-language

    said method comprising a first-step (202) of providing
    an additional-speech-data-corpus, said additional-
    speech-data-corpus comprising a collection of domain-
    specific-speech-data and/or dialect-specific-speech-
    data, and

    said method comprising a second-step (203) of
    reestimating acoustic-model-parameters of said base-
    speech-recognizer by a speaker-adaption-technique
    using said additional-speech-data-corpus.

13. Method for generating an adapted-speech-recognizer
     according claim 12

    said method comprising an optional third-step (204)
    for smoothing the reestimated acoustic-model-
    parameters.

14. Method for generating an adapted-speech-recognizer
     according claim 12 or 13

    said method comprising an optional fourth-step (205)

    for iterating said first-step by providing a modified
    additional-speech-data-corpus and

    for iterating said second and third-step based on said
    reestimated-acoustic-model-parameters or based on said
    base-acoustic-model-parameters.

15. Method for generating an adapted-speech-recognizer according claim 12 to 14

   wherein said acoustic-model is a Hidden-Markov-Model (HMM), and

   wherein said speaker-adaption-technique is the Maximum-A-Posteriori-adaption (MAP) or

   wherein said speaker-adaption-technique is the Maximum-Likelihood-Linear-Regression-adaption (MLLR), and

   wherein said third-step performing a Bayesian smoothing.

16. Method for generating an adapted-speech-recognizer according claim 12 to 15

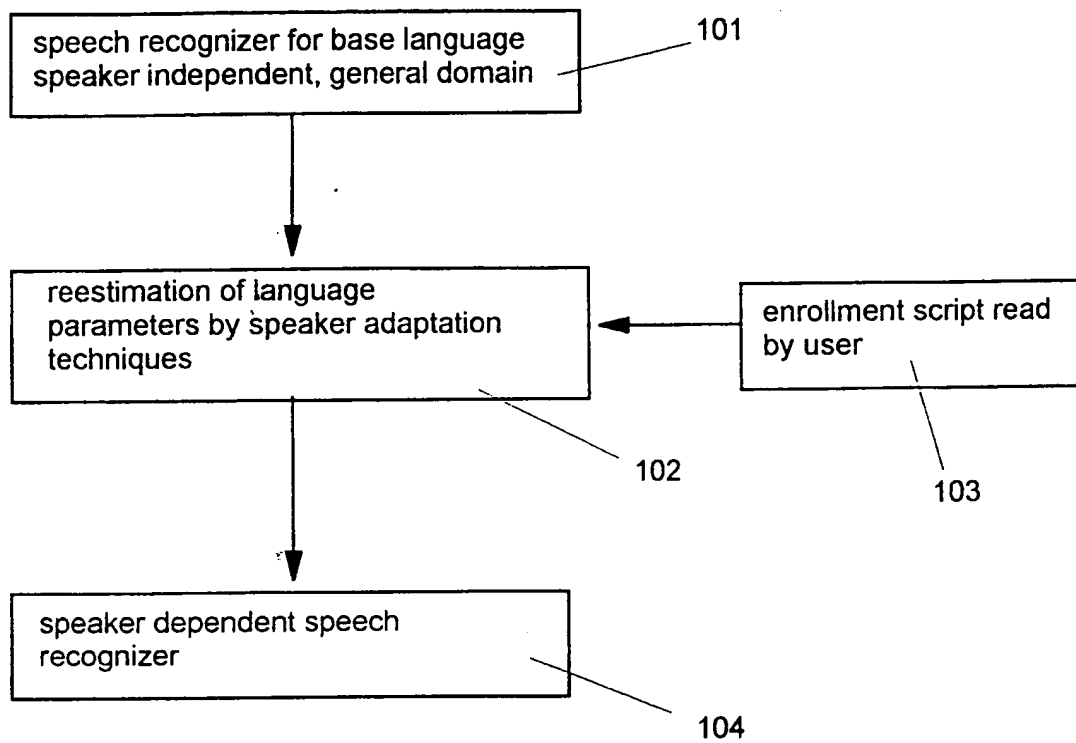   wherein said adapted-speech-recognizer being speaker independent.

```
┌─────────────────────────────────────┐
│ speech recognizer for base language  │──── 101
│ speaker independent, general domain  │
└─────────────────────────────────────┘
                  │
                  ▼
┌─────────────────────────────────────┐        ┌──────────────────────┐
│ reestimation of language             │◄───────│ enrollment script read │
│ parameters by speaker adaptation     │        │ by user                │
│ techniques                           │        └──────────────────────┘
└─────────────────────────────────────┘                    │
                  │            102                         103
                  ▼
┌─────────────────────────────────────┐
│ speaker dependent speech             │
│ recognizer                           │
└─────────────────────────────────────┘
                              104
```
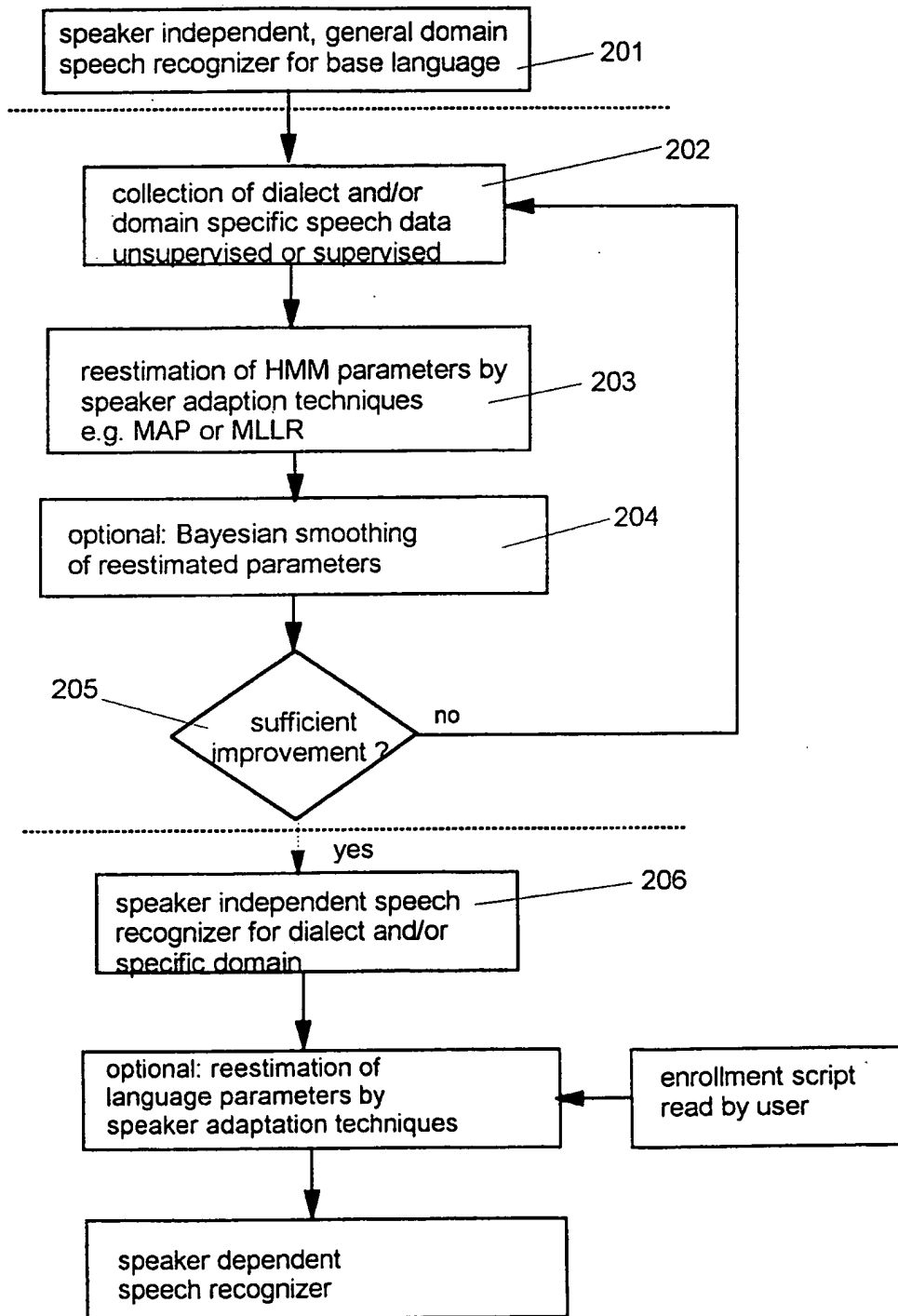
FIG. 1

FIG. 2

3 / 3

Test Speaker	VV	VV-S	smoothing 1	VV-G factor 50	k 500
German	1.00	1.04	1.30	1.26	1.19
Austrian	1.53	1.19	1.08	1.07	1.09

FIG. 3

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6   G10L5/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6   G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	DIAKOLOUKAS V ET AL: "Development of dialect-specific speech recognizers using adaptation methods" 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (CAT. NO.97CB36052), 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, MUNICH, GERMANY, 21-24 APRIL 1997, pages 1455-1458 vol.2, XP002111686 1997, Los Alamitos, CA, USA, IEEE Comput. Soc. Press, USAISBN: 0-8186-7919-0 paragraph '0002! paragraph '0003! --- -/--	1-3,12

☒ Further documents are listed in the continuation of box C.   ☐ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

& document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 August 1999	20/08/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Krembel, L

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	HSIAO-WUEN HON ET AL: "TOWARDS SPEECH RECOGNITION WITHOUT VOCABULARY-SPECIFIC TRAINING" PROCEEDINGS OF THE EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY (EUROSPEECH), PARIS, SEPT. 26 - 28, 1989, vol. 1, no. CONF. 1, 26 September 1989 (1989-09-26), pages 481-484, XP000209672 TUBACH J P;MARIANI J J paragraph '0004!	1,12
A	"BUILDING BASEFORMS FOR A NEW APPLICATION DOMAIN" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 4, 1 April 1993 (1993-04-01), pages 93-94, XP000364452 ISSN: 0018-8689 the whole document	1,12

4

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of KICHERER et al.

Application No. 10/084,099   Examiner:

Filed: February 27, 2002   Group Art Unit: 1772

For: SHAPED THERMAL INSULATION BODY

STATEMENT OF LACK OF DECEPTIVE INTENT

Commissioner for Patents
Box Fee Amendment
Washington, DC 20231

Sir:

I hereby declare that I did not willfully or with deceptive intent prevent my name from being included as co-inventor of above-identified patent application at the time of initial filing.

Dated:_____   _____
                                         Dr. Guenter Kratel